ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 3, May 2014

Issues of Web-Based Monitoring Implementation in Higher Education

Olga Cherednichenko, Olha Yanholenko, Iryna Liutenko, Abdugani Norbutaev

Abstract*— The paper is devoted to technology of web-based monitoring in higher education. It consists of three processes: data sources searching, data retrieval and indicators measurement. Data sources searching based on topic-focused web crawling provides a collection of potentially useful web pages. Data retrieval is aimed at clustering of collected web pages and extraction of raw data from them. Indicators measurement represents a measurement model for calculation of indicators' values. The suggested technology is implemented on the basis of multiagent approach.

***Index Terms*— comparator identification, measurement model, multiagent software, topic-focused web crawling, web-based monitoring, web pages clustering.**

I. INTRODUCTION

Nowadays educational services delivery complies with the rules of market economy. Every higher education establishment (HEE) tends to be more competitive and successful within the city, country or even worldwide. A good reputation among students, graduates and employers provides HEEs with a better contingent of entrants. The principles of competitive activities form the basis of university's development strategy. The comparison of achievements of HEEs is interesting not only to potential students, but also to university management. In order to regulate the problem of HEEs comparison different university rankings are developed by non-governmental organizations [1].

The comprehensive university's estimate, which defines its place in the ranking, is influenced by many factors. The set of such factors is determined by organization that creates the ranking. Usually these factors include academic reputation, reputation among employers, level of students' satisfaction, quality of researches and teaching, quality of resources, cultural impact and level of incomes and expenses [2]. Analyzing the HEE's place in the ranking its management can investigate the impact of each factor on the comprehensive estimate. Such analysis can help a university to improve its strategy or the policy in some distinct direction.

One of the advantages of such comprehensive assessment provided by rankings is that the obtained estimates are external in relation to universities. Another advantage is that the rankings are constructed at some intervals, for instance, annually. So it is possible to trace the HEE's position within time and make conclusions about the results of management policy.

Modern organizations, as well as universities, rely not only on the estimates of external experts, but also make the self-assessment for the purpose of their market position improvement. The monitoring methods allow collecting the data for indicators estimation [3]. There are different methods of data collection. Our work is devoted to web-based methods of data collection that consider the web as a source of data.

We discovered that a lot of data on the web reflects the outcomes of HEEs activities. Such data can be gathered from corporative web sites, social networks, rankings, news, blogs, information portals which publish reports about research and educational activities, courseware, and so on. Methods used for such data analysis are based on data mining and statistical analysis [4]. Comparing to traditional monitoring the web-based monitoring gives some advantages (Table 1).

The goal of this work is to model the process of web-based monitoring in HEEs in order to design the software solution for its automation.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 3, May 2014

Table 1. Comparison of traditional and web-based monitoring

Comparison criterion	Traditional monitoring	Web-based monitoring
Sources of data	Official documentation, reports, surveys, interviews	Relevant web pages
Methods of data collection	Reviews, questionnaires, interviewing	Web mining, data mining
Possibility of automation	Partially automated	Totally automated
Cost-effectiveness	High expenses	Low expenses
Experts involvement	Expert judgments	Ontology

II. PROBLEM STATEMENT

The technology of web-based monitoring is the subject of research of many authors. It has found applications in different domains. For example, web-based monitoring is the basis of competitive intelligence [5]. It is quite popular in economical area, for instance, it can be used for the formation of the price strategy of the enterprise [6]. Our previous researches are devoted to monitoring and evaluation problems in higher education [7], [8]. We obtained the framework of quality monitoring. Its goal is to provide values of indicators. The key elements of monitoring framework describe what data should be gathered, where it can be found and how it can be extracted. The data and its sources have to be defined by experts. In the case of web-based monitoring this also holds true. An expert should specify the kind of data and where it can be found, i.e. web sites together with its position on web page. The way how the data must be collected is determined by methods of web mining that are realized in particular information system (IS).

As a rule, the automation of web-based monitoring is implemented by means of the functionality provided by well-known search engines. In our work we deal with the problem of topic-focused web search. This means that the main criterion of web page's relevance for us is its similarity to some topic. The popularity of a web page among users, its link characteristics are not so essential. Also for the purposes of monitoring it is extremely important to have the actual data. Therefore the time of web page crawling, indexing and estimation must meet the requirements of timeliness. The existing search engines are not suitable for the purpose of topical web search. Moreover they can't be used for indicators' measurement. After the possible sources of data are defined we need to find the values of indicators based on the data from web pages. Such functionality together with topical search is the basis of the monitoring system which we are going to implement.

III. THE GENERAL TECHNOLOGY OF WEB-BASED MONITORING

The key processes of web-based monitoring include data sources searching, data retrieval and indicators measurement (fig. 1). In such system an expert is a person who brings the necessary knowledge. An expert must define how the data from the web can be transformed into indicators of university's functioning. When an expert formulates indicators that reflect the outcomes, he thinks over the data from the web which can be used to estimate them. Such knowledge is formalized in the ontology. It contains the information about indicators, their topics of search, necessary terms and their synonyms, places where they can be found, metadata of web pages that are to be investigated to find required data.

The goal of data sources searching process is to collect web pages which contain the necessary data for indicators measurement. In our work such process is implemented in the form of topic-focused web crawling. The topic of search is defined in ontology. The result of this process is the collection of web pages.

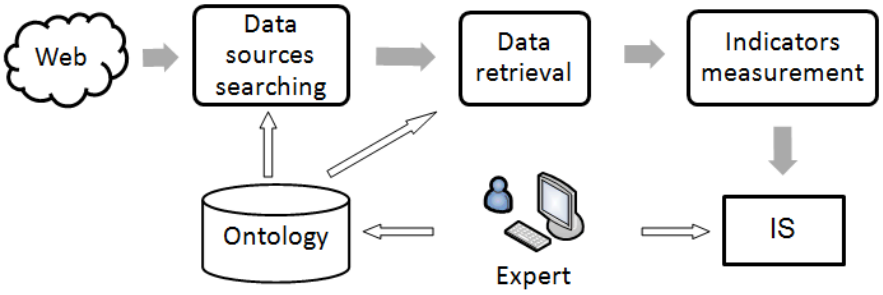

Fig. 1. Web-based monitoring

Data retrieval process is aimed at extraction of raw data necessary for indicators calculation. Collected web pages

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 3, May 2014

have to be grouped in order to find the duplicates and define web pages associated with each indicator. In our research we implement this process in the form of topical clustering. After the duplicates are eliminated, the web pages must be parsed for raw data extraction.

Indicators measurement is implemented based on the measurement model which can be either statistical in the case of a big volume of available data or evidentiary in the opposite case. Our research considers statistical models, for instance the models of Item Response Theory [9]. As a result the raw data is transformed into values of indicators. Let's consider the mentioned processes in more details. We suggest to use a multiagent software paradigm for monitoring system development. This gives us all advantages of distributed open software. The multiagent approach allows creating efficient, scalable and portable solutions that incorporate the intelligent properties of software agents.

In this work we suggest the following multiagent realization of data sources searching process (fig. 2). The agents of three types are involved in the process. The agent of the first type (A1) is a crawler itself. It fetches a Parent URL, parses, extracts the out-links and estimates whether the web page is promising for the further search. If the estimate is positive, the links are sent to agent of the second type (A2) which is responsible for adding new Child URLs to the database. The agent of the third type (A3) selects the next Parent URLs from the database and assign A1 agents for their processing.

The estimate of a web page is done by A1 based on the method of comparator identification [10]. It allows to match the topic of a web page and the topic of search based on the values of subject variables. This method models the estimation process as a human intelligent activity, since a human looking through a web page can easily determine whether it is useful for him or not.

The following subject variables are considered: terms in page's title – $t = \overline{1,m}$, terms in page's keywords – $k = \overline{1,n}$, terms in page's hyperlinks text – $h = \overline{1,p}$, and a topic of search – q. A searching topic q_i is associated with a collection of terms from a web page's title, keywords and hyperlinks text. This can be expressed through the predicates: $P_t(t,q_i) = t_1 \vee t_2 \vee ... \vee t_m$, $P_k(k,q_i) = k_1 \vee k_2 \vee ... \vee k_n$ and $P_h(h,q_i) = h_1 \vee h_2 \vee ... \vee h_p$. According to these predicates a web page belongs to a topic against a title, keywords and hyperlinks text if at least one term appears in the corresponding place.

The web page's topic is defined by a composite predicate: $S = P(t,k,h,q) = P_t(t,q) \vee P_k(k,q) \vee P_h(h,q)$. The estimate of web page's perspective is expressed by the predicate $S(q^*)$, where q^* is a topic of search. The result of data sources searching process is the collection of web pages that have to be processed next.

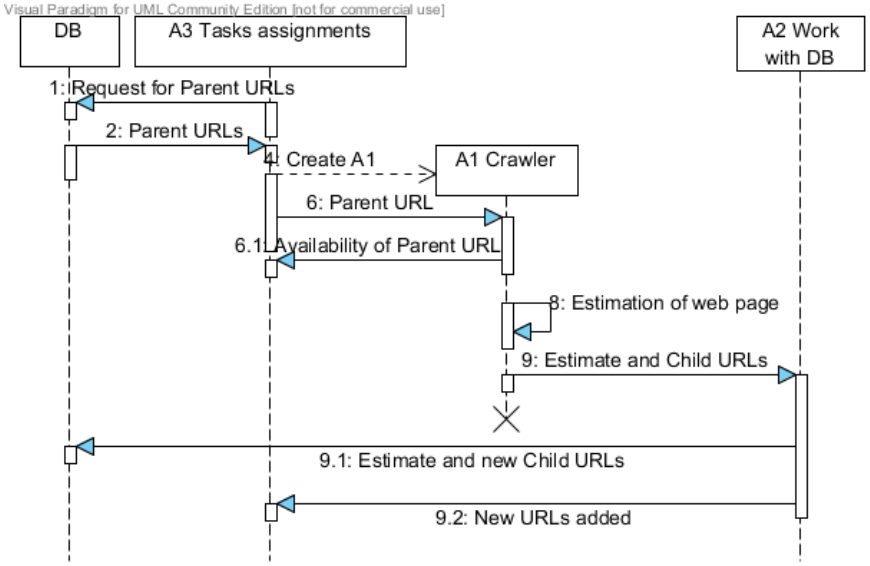

Fig. 2. Data sources searching

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 3, May 2014

As it was mentioned, data retrieval process includes clustering of web pages and documents parsing. In order to provide the clustering procedure we suggest to use similarity measure of web pages as clustering criteria. The similarity measure is defined by web page's descriptors, for example metadata elements. To estimate the similarity measure of two documents we use the method of comparator identification (fig. 3). As an input we have documents' descriptors. Analyzing them we can define the topics of documents and then compare them. We suggest to determine the similarity measure by the formula:

$$m(d_1, d_2) = \sum a_i m_i(d_1, d_2),$$

where $m_i(d_1, d_2)$ is a similarity measure of documents d_1 and d_2 subject to *i*-th descriptor;

a_i is a weight coefficient of *i*-th descriptor, $\sum a_i = 1$.

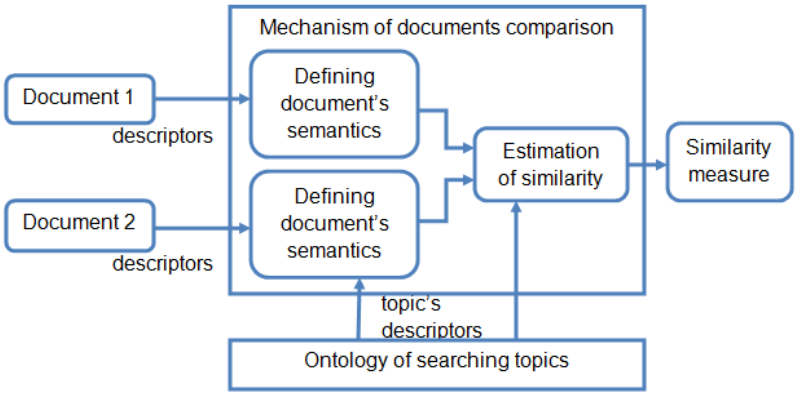

Fig. 3. Similarity measure estimation based on comparator identification

The similarity measure $m_i(d_1, d_2)$ subject to *i*-th descriptor is defined according to some rule. For example, such rule may state that if two documents have at least one common term in their titles, then $m_i(d_1, d_2) = 1$, otherwise $m_i(d_1, d_2) = 0$. Introduction of weight coefficients allows the coincidence in particular descriptors to make more important.

In order to implement such methodology we assign an agent to each web page. We suppose that agents may form coalitions which will represent clusters of similar web documents. Web page's metadata have been extracted by a crawling agent A1 and stored in the database. Every agent must communicate with all other *N* agents (fig. 4). During the communication agents must exchange with their descriptors. Every agent *j* calculates the similarity measure $m_\gamma(d_j, d_\gamma)$, $\gamma = \overline{1, N}$, $\gamma \neq j$ between itself and all other agents. A proposal of agent *j* to form a coalition is accepted by agent γ, if from its point of view agent *j* also has a maximal similarity measure. In the case if the coalition isn't formed, agent *j* sends the proposal to the next best agent from the list.

After the first round of communication some coalitions are formed. In the next rounds the procedure of descriptors exchange and similarity measure estimation is done only for the new agents that appear. The agents, who are not in the coalition yet, try to form the coalitions with single agents or with existing coalitions.

Finally we suggest to implement indicators measurement process with the help of Rasch Model, which is one of the models of Item Response theory [9]. For this purpose we need to form the assessment matrix $N \times M$ with elements $x_{ij} = \{0,1\}$, where $i \in \overline{1, N}$ is the term from the ontology and $j \in \overline{1, M}$ is the targeted web page. x_{ij} is equal to 1, if the *i*-th term is present on *j*-th web page, otherwise it equals 0. According to Rasch model the indicator's value is defined from the following dependency:

$$P(x_{ij} = 1 \mid \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)},$$

where $P(x_{ij})$ is probability of presence of *i*-th term at *j*-th web page;

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 3, May 2014

θ_i is a value of *i*-th indicator associated with *i*-th term;

β_j is difficulty of *j*-th web page.

This procedure is implemented by assessment agent. Another agent must estimate the reliability of obtained results which is expressed through KR20 coefficient. As a result we obtain the values of specified indicators.

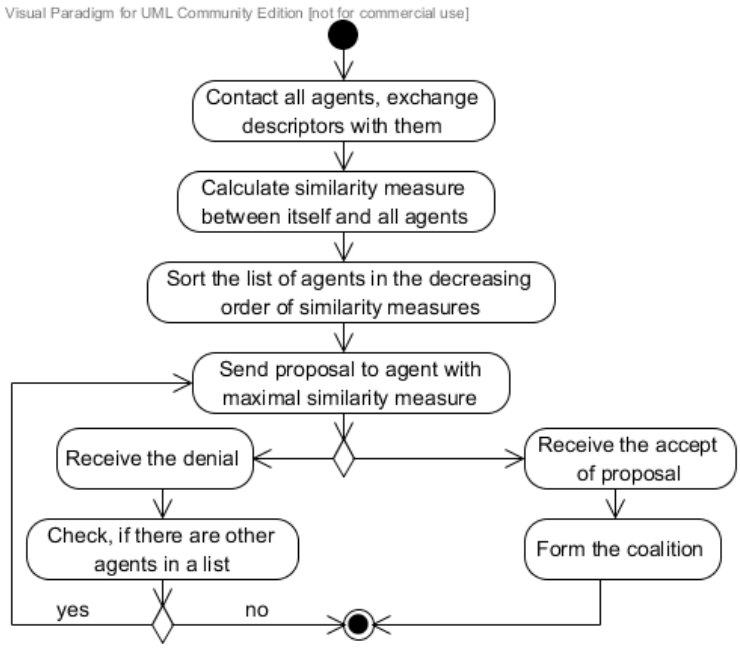

Fig. 4. A single agent's behavior in the first round of communication

IV. CONCLUSIONS

The web-based monitoring is an alternative way of self-assessment for modern universities. The suggested technology of web-based monitoring covers three key processes: data sources searching, data retrieval and indicators measurement. We defined the models that underlie each of these processes. Data sources searching is realized through topic-focused web crawling. Data retrieval requires web pages clustering based on comparator identification. Indicators measurement is supported by statistical models. The suggested multiagent implementation of the technology of web-based monitoring makes the system more flexible and scalable subject to adding new resources.

REFERENCES

[1] QS World university rankings. Retrieved March 02, 2014 from http://www.topuniversities.com/university-rankings/world-university-rankings.

[2] QS Stars methodology. Retrieved March 03, 2014 from http://www.iu.qs.com/services/qs-stars/qs-stars-methodology.

[3] J.Z. Kusek, R.C. Rist, "Ten steps to a results-based monitoring and evaluation system: a handbook for development practitioners", The World Bank, Washington, DC, 2004.

[4] C. Manning., P. Raghavan, H. Schutze, "An introduction to information retrieval", Cambridge University Press, Cambridge, 2009.

[5] D.V. Lande, "Searching of knowledge on the Internet", Moscow, Publishing house "Williams", 2005.

[6] I.A. Cherenkov, S.V. Orekhov, "Data retrieval from textual news on the example of polymer market", Systems of information processing, vol. 9(107), pp. 224-228, 2012.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 3, May 2014

[7] O. Cherednichenko, O. Yangolenko, "Towards Quality Monitoring and Evaluation Methodology: Higher Education Case-Study", In: H.C. Mayr et al. (Eds.): UNISCON 2012, LNBIP, vol. 137, pp. 120-127, 2013.

[8] O. Cherednichenko, O. Yanholenko, I. Liutenko, O. Iakovleva, "Monitoring and Evaluation Problems in Higher Education: Comprehensive Assessment Framework Development", Proc. of the 5-th International Conference on Computer Supported Education CSEDU 2013, SCITEPRESS, pp. 455-460, 2013.

[9] B. Wright, M. Stone, "Measurement Essentials (2nd ed.)", Wilmington, Wide Range Inc, 1999.

[10] M.F. Bondarenko, U.P. Shabanov-Kushnarenko, "Theory of intelligence: a Handbook", Kharkiv, SMIT Company, 2006.

AUTHOR BIOGRAPHY

Olga Cherednichenko is an associate professor of National Technical University "Kharkiv Polytechnic Institute" on the department of Computer-Aided Management Systems. Her research interest is management of social and economical systems, in particular higher education quality management system. Her recent publications are devoted to the problems of web-based monitoring and evaluation of the results of universities functioning and to the issues of comprehensive assessment of resources quality in higher education. She promotes the development of distributed software of intelligent systems on the basis of multiagent approach. She obtained her M. Sc. And PhD degrees in National Technical University "Kharkiv Polytechnic Institute".

Olha Yanholenko is a post-graduate student of National Technical University "Kharkiv Polytechnic Institute" on the department of Computer-Aided Management Systems. Her primary research interests include web-based monitoring and evaluation of the results of university's scientific activities. Her publications are devoted to multiagent implementation of web crawling and web mining intelligent processes, and to statistical processing of data collected from the web. She obtained her M. Sc. degree in National Technical University "Kharkiv Polytechnic Institute".

Iryna Liutenko is a senior lecturer of National Technical University "Kharkiv Polytechnic Institute" on the department of Computer-Aided Management Systems. Her primary research interest is resources quality management in the system of higher education. Her publications are devoted to the problems of comprehensive assessment of resources quality and to the automation of licensing and accreditation procedures in the university. She obtained her M. Sc. degree in National Technical University "Kharkiv Polytechnic Institute".

Abdugani Norbutaev is a master student of National Technical University "Kharkiv Polytechnic Institute" on the department of Computer-Aided Management Systems. His research interest is development of distributed software of intelligent systems.